

Automating work in Galaxy

Tom Doak

Le-Shin Wu

Carrie Ganote

National Center for Genome Analysis Support

August 12, 2015



INDIANA UNIVERSITY



INDIANA UNIVERSITY



Automation Overview

There are two main approaches to automating work in Galaxy:

- Running one tool on lots of inputs
- Creating and running workflows





Multiple inputs in one tool

Short read data from your current history:  



5: FASTQ Quality Trimmer on data 2

Some tools, such as FastQC, support multiple inputs. Note that these will run as separate, unrelated jobs!

Short read data from your current history:  

1: TB_1.fq
2: TB_2.fq
4: FASTQ Quality Trimmer on data 1
5: FASTQ Quality Trimmer on data 2

Click on “Run tool in parallel..”

Short read data from your current history:  

1: TB_1.fq
2: TB_2.fq
4: FASTQ Quality Trimmer on data 1
5: FASTQ Quality Trimmer on data 2

Use the shift and control keys to toggle or select many inputs at once.



Introduction to workflows

Tools

search tools

Import Data

Data Manipulation

Quality Control

De novo Assembly

Mapping and Alignments

Run Blast+

Annotation

Statistics

Variants

Clustering/Phylogeny

Visualization

NGS: Mapping

Workflows

■ All workflows

Your workflows

[Switch to workflow management view](#)

Name	# of Steps
Transcriptome Assembly Workflow	12
Testing multiple dsets	2
Unnamed workflow	1
Workflow constructed from history	18
Workflow constructed from history	17
Workflow constructed from history 'Tuxedo Suite	15
test2	28
Workflow constructed from history	27
Workflow constructed from history	5
Workflow constructed from history	14
Gene enhancer region analysis (imported from uploaded file)	7
imported: Workflow constructed from history 'Galaxy 101'	5
imported: Workflow constructed from history 'Galaxy 101'	5
imported: Workflow constructed from history 'Galaxy 101'	5
imported: Workflow constructed from history 'Galaxy 101'	5
Workflow constructed from history 'Galaxy 101'	5
Workflow constructed from history 'imported: new data set' (imported from uploaded file)	220

Workflows shared with you by others

No workflows have been shared with you.

Workflows encapsulate an entire histories worth of work. They can be created by:

- Building from scratch
- Extracting from history
- Importing from shared workflow



Building workflows

Your workflows

Switch to workflow management view

Your workflows

Create new workflow

Upload or import workflow

Create New Workflow

Workflow Name:

Yet another test

Workflow Annotation:

Something descriptive

A description of the workflow; annotation is shown alongside shared or published workflows.

Create

Create a new workflow and
annotate it with something
you will remember a year
from now!



Building workflows

Each step is created by choosing tools to build the pipeline. Here is an example input with a dataset collection.

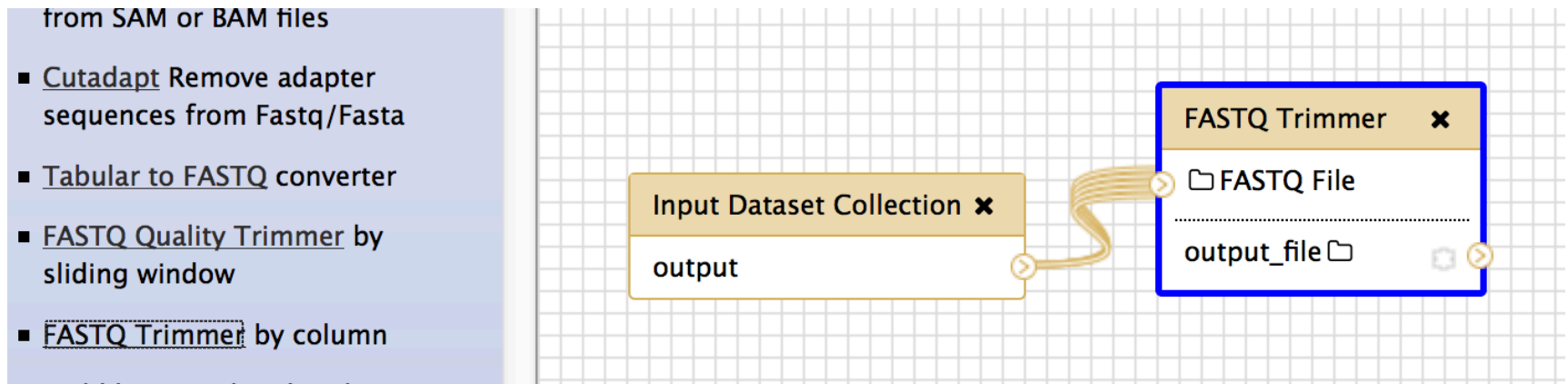
The screenshot displays the NCGAS workflow builder interface, which is divided into three main sections: Tools, Workflow Canvas, and Details.

- Tools:** A sidebar on the left containing a search bar and a list of tool categories: Import Data, Data Manipulation, Quality Control, De novo Assembly, Mapping and Alignments, Run Blast+, Annotation, Statistics, Variants, Clustering/Phylogeny, Visualization, NGS: Mapping, Data Manager Tools, Workflow control, Inputs, and Outputs. The 'Inputs' section is currently expanded, showing 'Input dataset' and 'Input dataset collection'.
- Workflow Canvas:** A central grid area where workflow steps are placed. A blue-bordered box labeled 'Input Dataset Collection' is being added to the canvas. The box has a title bar with a close button (x) and a label 'output' with a plus icon.
- Details:** A sidebar on the right showing the configuration for the selected tool. It includes fields for 'Name' (set to 'All paired end reads from experiment'), 'Collection Type' (set to 'list'), and a section for 'Annotation / Notes' with a text area and a note: 'Add an annotation or notes to this step; annotations are available when a workflow is viewed.'

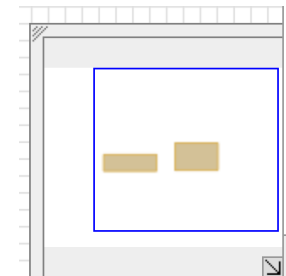


Building workflows

Outputs from one tool can be clicked and dragged into the input of another tool. This process can be extended until the entire pipeline is created.



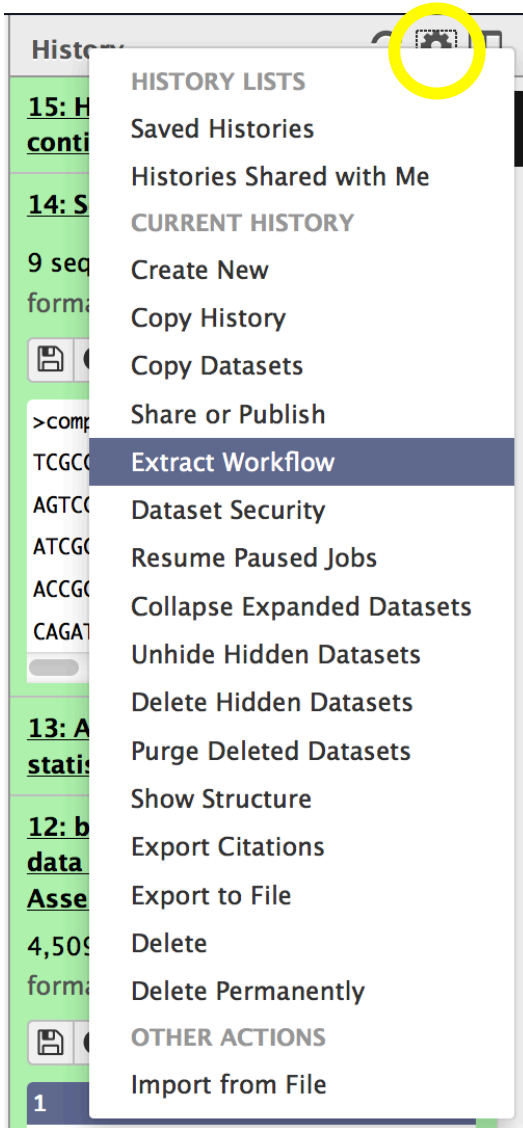
The bottom right side has an overview panel to allow the entire board to be scrolled by clicking and dragging.





Extracting workflows

An entire history can be converted into a workflow for easy sharing and to reproduce work. Make sure you are in the history that you want to extract from.





Extracting workflows

As always, a descriptive name will help in the long run!

Create Workflow when done

Input files are chosen later.

You can skip steps you don't need

The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

Workflow name

Workflow constructed from history 'Test2015'

Create Workflow

Check all

Uncheck all

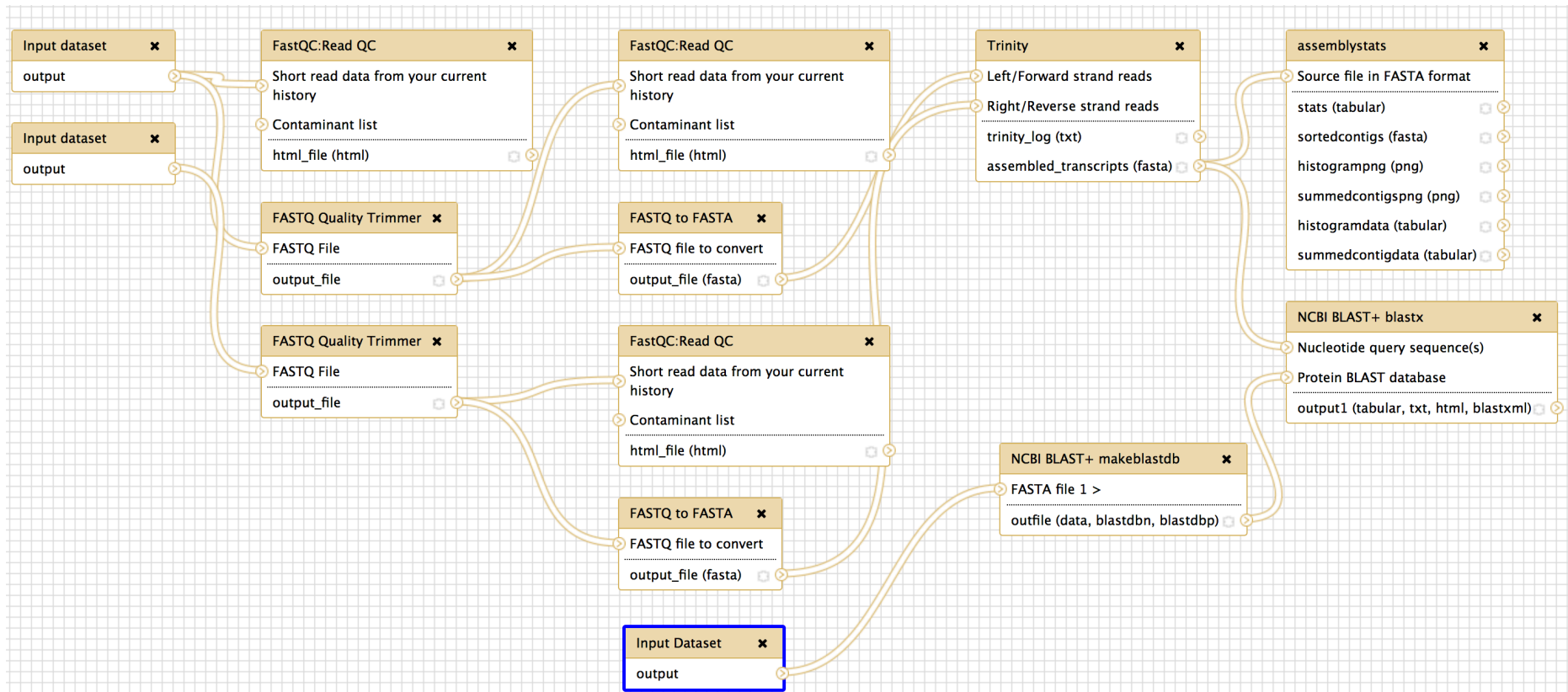
Tool

History items created

Unknown <i>This tool cannot be used in workflows</i>	▶	1: TB_1.fq <input checked="" type="checkbox"/> Treat as input dataset
Unknown <i>This tool cannot be used in workflows</i>	▶	2: TB_2.fq <input checked="" type="checkbox"/> Treat as input dataset
FastQC:Read QC <input type="checkbox"/> Include "FastQC:Read QC" in workflow	▶	3: FastQC_TB_1.fq.html
FASTQ Quality Trimmer <input checked="" type="checkbox"/> Include "FASTQ Quality Trimmer" in workflow	▶	4: FASTQ Quality Trimmer on data 1
FASTQ Quality Trimmer <input checked="" type="checkbox"/> Include "FASTQ Quality Trimmer" in workflow	▶	5: FASTQ Quality Trimmer on data 2



Extracting workflows





Importing workflows

Galaxy / at IU

Published Workflows

search name, annotation, owner, and tags

[Advanced Search](#)

Name

Transcriptome Assembly Workflow

Workflow constructed from

Import

Save as File

Shared Data ▾

Visualization ▾

Data Libraries

Data Libraries Beta

Published Histories

Published Workflows

Published Visualizations

Published Pages

Sharing workflows makes it easy for others to reproduce your work and for you to run new inputs through the same procedure



Running workflows

The inputs are the only thing you need to set up in order to run the workflow.

For a cleaner history, send results to a new history:

Running workflow "Transcriptome Assembly Workflow"

This workflow takes 2 fastq files as input and trims, assembles, and blasts the data.

Step 1: Input dataset

Left Reads Fastq File

- 1: TB_1.fq
- 1: TB_1.fq
- 2: TB_2.fq
- 4: FASTQ Quality Trimmer on data 1
- 5: FASTQ Quality Trimmer on data 2

Right Reads Fastq File

- 1: TB_1.fq
- type to filter

Step 3: Input dataset

Use tb prot file

Fasta genome, protein sequences

- 94: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000195955.2_ASM19595v2/GCF_000195955.2_ASM
- type to filter

Step 4: FastQC:Read QC (version 0.52)

Step 5: FASTQ Quality Trimmer (version 1.0.0)

Step 6: FASTQ Quality Trimmer (version 1.0.0)

Step 7: NCBI BLAST+ makeblastdb (version 0.0.22)

Step 8: FastQC:Read QC (version 0.52)

Step 9: FASTQ to FASTA (version 1.0.0)

Step 10: FastQC:Read QC (version 0.52)

Step 11: FASTQ to FASTA (version 1.0.0)

Step 12: Trinity (version 0.0.1)

Step 13: assemblystats (version 1.0.1)

Step 14: NCBI BLAST+ blastx (version 0.0.22)

☐ Send results to a new history

Run workflow



Sharing workflows

Your workflows

Switch to workflow management view

Your workflows

Name

Transcriptome Assembly Workflow

Edit

Run

Share or Publish

Download or Export

Copy

Rename

View

Delete

Sharing workflows makes it easy for others to reproduce your work and for you to run new inputs through the same procedure

Share or Publish Workflow 'Transcriptome Assembly Workflow'

Make Workflow Accessible via Link and Publish It

This workflow is currently accessible via link and published. Anyone can view and import this workflow by visiting the following URL:

<https://galaxy.iu.edu/galaxy-upgrade/u/cganote/w/transcriptome-assembly-workflow>

This workflow is publicly listed and searchable in Galaxy's [Published Workflows](#) section.

You can:

Unpublish Workflow

Removes this workflow from Galaxy's [Published Workflows](#) section so that it is not publicly listed or searchable.

Disable Access to Workflow via Link and Unpublish

Disables this workflow's link so that it is not accessible and removes workflow from Galaxy's [Published Workflows](#) section.

Share Workflow with Individual Users

You have not shared this workflow with any users.

Share with a user

[Back to Workflows List](#)



Future directions

There are other methods in development that will add even more flexibility to running massive numbers of jobs and increasing throughput:

- API calls for advanced use
- Dataset collections
- Nested workflows



INDIANA UNIVERSITY

Fin

Thanks for watching!
Questions and comments:
Email help@ncgas.org